University of Montpellier Year 2026-2027
HAX912X - Generalized Linear Model / Machine Learning

# Additional work on Generalized Linear Models

**Exercise 1**

Consider a $n$-sample $(y_1, x_1), \ldots, (y_n, x_n)$ of the pair $(y, x)$ where $y$ is a random variable with values in $\{0, 1\}$ and $x$ is a fixed variable also taking the values $\{0, 1\}$. It is assumed that there exists a latent variable $y_i^*$ such that $y_i^* | x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, 1)$ and

$$y_i = \begin{cases} 1 & \text{si} \quad y_i^* \leq 0 \\ 0 & \text{si} \quad y_i^* > 0 \end{cases} .$$

The distribution function of the standard gaussian distribution is denoted by $\Phi(\cdot)$ ($\Phi(u) = \mathbb{P}(\mathcal{N}(0,1) \leq u)$).

**1** Show that this is a Generalised Linear Model. Following the notations of the course for exponential families, explicit the parameters $\phi$, $\theta$, the functions $b$, $c$, $\gamma$ and the regressors to be considered.

**2** Has the canonical link been chosen?

**3** We have 100 observations $(y_i, x_i)$ described in the following contingency table

| | $y = 0$ | $y = 1$ |
|---|---|---|
| $x = 0$ | 10 | 30 |
| $x = 1$ | 20 | 40 |

.

Let $\alpha_0 = \Phi(\beta_0)$ and $\alpha_1 = \Phi(\beta_0 + \beta_1)$ give the log-likelihood as a function of $\beta_0$ and $\beta_1$.

**4** Calculate the maximum likelihood estimator for $(\alpha_0, \alpha_1)$.

**Exercise 2**

Consider $n$ independent random variables $y_1, \ldots, y_n$ such that $y_i \sim \mathcal{N}\left(\exp(\beta_0 + \beta_1 x_i), \sigma^2\right)$.

**1** Show that this is a Generalized Linear Model. Following the notations of the course for exponential families with nuisance parameter, explicit $\phi$, $\theta$, the functions $\gamma$ and $b$ and the regressors to be considered.

**2** Show that the canonical link has not been chosen.

**3** Assuming in the following that $\sigma^2 = 1$, give the log-likelihood and the likelihood equations. Can we calculate the analytical expressions for the maximum likelihood estimators of $\beta_0$ and $\beta_1$?

# Corrections

## Correction of Exercise 1

Let

$$\eta_i = \beta_0 + \beta_1 x_i.$$

Since $y_i^* \mid x_i \sim \mathcal{N}(\eta_i, 1)$ and $y_i = 1$ if and only if $y_i^* \leq 0$,

$$\mathbb{P}(y_i = 1 \mid x_i) = \mathbb{P}(y_i^* \leq 0 \mid x_i) = \Phi(-\eta_i) = 1 - \Phi(\eta_i).$$

Thus, if $\mu_i = \mathbb{E}(y_i \mid x_i)$,

$$\mu_i = \Phi(-\eta_i), \qquad \eta_i = -\Phi^{-1}(\mu_i).$$

**1** Conditionally on $x_i$, $y_i$ follows a Bernoulli distribution with parameter $\mu_i$. Its density can be written

$$f(y_i; \mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left\{ y_i \log \frac{\mu_i}{1-\mu_i} + \log(1-\mu_i) \right\}.$$

This is an exponential family. With the usual notation

$$f(y; \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\},$$

we may take

$$\phi = 1, \qquad \theta = \log \frac{\mu}{1-\mu}, \qquad b(\theta) = \log(1 + e^\theta), \qquad c(y, \phi) = 0.$$

Moreover,

$$b'(\theta) = \frac{e^\theta}{1+e^\theta} = \mu, \qquad b''(\theta) = \mu(1-\mu).$$

The link function is

$$\gamma(\mu) = -\Phi^{-1}(\mu),$$

because

$$\gamma(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i.$$

The regressor vector is therefore $(1, x_i)$ and the parameter vector is $\beta = (\beta_0, \beta_1)^T$. Hence the model is a binomial GLM with a probit link (with a minus sign due to the convention $y_i = 1$ when $y_i^* \leq 0$).

**2** For a Bernoulli distribution, the canonical link is the logit link $\log \frac{\mu}{1-\mu}$. Here the link is $\gamma(\mu) = -\Phi^{-1}(\mu)$, so the canonical link has not been chosen.

**3** Put

$$\alpha_0 = \Phi(\beta_0), \qquad \alpha_1 = \Phi(\beta_0 + \beta_1).$$

Because of the convention used in the statement,

$$\mathbb{P}(y = 1 \mid x = 0) = 1 - \alpha_0, \qquad \mathbb{P}(y = 0 \mid x = 0) = \alpha_0,$$

and

$$\mathbb{P}(y = 1 \mid x = 1) = 1 - \alpha_1, \qquad \mathbb{P}(y = 0 \mid x = 1) = \alpha_1.$$

The likelihood is therefore

$$\alpha_0^{10}(1-\alpha_0)^{30}\alpha_1^{20}(1-\alpha_1)^{40} = \Phi(\beta_0)^{10}[1-\Phi(\beta_0)]^{30}\Phi(\beta_0+\beta_1)^{20}[1-\Phi(\beta_0+\beta_1)]^{40}.$$

Hence the log-likelihood is

$$10\log\Phi(\beta_0) + 30\log[1-\Phi(\beta_0)] + 20\log\Phi(\beta_0+\beta_1) + 40\log[1-\Phi(\beta_0+\beta_1)]$$

$$= 10\log\alpha_0 + 30\log(1-\alpha_0) + 20\log\alpha_1 + 40\log(1-\alpha_1).$$

**4** The two terms involving $\alpha_0$ and $\alpha_1$ can be maximized separately. The partial derivative of the log-likelihood with respect to $\alpha_0$ gives

$$\frac{10}{\alpha_0} - \frac{30}{1-\alpha_0} = 0,$$

so

$$10(1-\alpha_0) = 30\alpha_0 \quad \Longrightarrow \quad \widehat{\alpha}_0 = \frac{10}{40} = \frac{1}{4}.$$

Similarly for $\alpha_1$,

$$\frac{20}{\alpha_1} - \frac{40}{1-\alpha_1} = 0,$$

which gives

$$\widehat{\alpha}_1 = \frac{20}{60} = \frac{1}{3}.$$

Thus

$$\widehat{\alpha}_0 = \frac{1}{4}, \qquad \widehat{\alpha}_1 = \frac{1}{3}.$$

If desired, the corresponding estimators of the original parameters are

$$\widehat{\beta}_0 = \Phi^{-1}\left(\frac{1}{4}\right), \qquad \widehat{\beta}_1 = \Phi^{-1}\left(\frac{1}{3}\right) - \Phi^{-1}\left(\frac{1}{4}\right).$$

Numerically, $\widehat{\beta}_0 \simeq -0.6745$ and $\widehat{\beta}_1 \simeq 0.2438$.

## Correction of Exercise 2

Write

$$\mu_i = \exp(\beta_0 + \beta_1 x_i) > 0.$$

We have $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

**1** The Gaussian density is

$$f(y_i;\mu_i,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right].$$

Expanding the square,

$$f(y_i;\mu_i,\sigma^2) = \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}.$$

Thus, in exponential-family notation,

$$\phi = \sigma^2, \qquad \theta_i = \mu_i, \qquad b(\theta) = \frac{\theta^2}{2}.$$

The nuisance term may be written

$$c(y, \phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi).$$

Since

$$\mu_i = e^{\beta_0 + \beta_1 x_i},$$

the link function is

$$\gamma(\mu) = \log \mu,$$

for which

$$\gamma(\mu_i) = \beta_0 + \beta_1 x_i.$$

The regressor vector is $(1, x_i)$. Therefore this is a Gaussian GLM with a log link.

**2** For the Gaussian family with known variance, the canonical parameter is $\theta = \mu$. Consequently the canonical link is the identity link. The model instead uses $\gamma(\mu) = \log \mu$. Therefore the canonical link has not been chosen.

**3** Assume $\sigma^2 = 1$. Since

$$\mu_i = e^{\beta_0 + \beta_1 x_i},$$

the log-likelihood is

$$\ell(\beta_0, \beta_1) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left[y_i - e^{\beta_0 + \beta_1 x_i}\right]^2.$$

Let $\eta_i = \beta_0 + \beta_1 x_i$ and $\mu_i = e^{\eta_i}$. Since

$$\frac{\partial \mu_i}{\partial \beta_0} = e^{\beta_0 + \beta_1 x_i} \qquad \frac{\partial \mu_i}{\partial \beta_1} = x_i e^{\beta_0 + \beta_1 x_i}$$

the score equations are

$$\sum_{i=1}^{n}\left[y_i - e^{\beta_0 + \beta_1 x_i}\right] e^{\beta_0 + \beta_1 x_i} = 0$$

$$\sum_{i=1}^{n} x_i \left[y_i - e^{\beta_0 + \beta_1 x_i}\right] e^{\beta_0 + \beta_1 x_i} = 0.$$

In general these equations are nonlinear in $(\beta_0, \beta_1)$, so there is no simple closed-form expression for the maximum likelihood estimators.

**Remark** The log link imposes $\mu_i > 0$, although observations from a Gaussian distribution may themselves be negative. This is not a contradiction: the restriction concerns the conditional mean, not the support of $y_i$.